R E P O R T R E S U M E S

ED 002 777 24

RESTRICTED GENERALIZATION, BIAS, AND LOSS OF POWER THAT MAY RESULT FROM MATCHING GROUPS.

BY- STANLEY, JULIAN C. BEEMAN, ELLEN Y.
WISCONSIN UNIV., MADISON

REPORT NUMBER CRP-149 PUB DATE 27 SEP 57
EDRS PRICE MF-$0.09 HC-$1.32 33P.

DESCRIPTORS- *MENTAL RETARDATION, METHODOLOGY, *PSYCHOLOGICAL STUDIES, *LITERATURE, *RESEARCH METHODOLOGY, MADISON, WISCONSIN

THIS STUDY CRITICALLY EVALUATED MAJOR PUBLISHED STUDIES DEALING WITH PSYCHOLOGICAL AND EDUCATIONAL ASPECTS OF MENTAL RETARDATION AND PREPARED AN EXTENSIVE, DETAILED ANALYSIS OF RESEARCH METHODOLOGY ON THIS SUBJECT. THIS WAS A LIBRARY RESEARCH PROJECT IN WHICH A GREAT BODY OF LITERATURE WAS SURVEYED, ESPECIALLY PERIODICALS. SELECTED CLUSTERS OF STUDIES AND SINGLE STUDIES ILLUSTRATING THE METHODOLOGICAL POINTS WERE EMPHASIZED. A FEW RESEARCHERS IN MENTAL DEFICIENCY ARE PERFORMING THEORETICALLY ORIENTED EXPERIMENTS, UTILIZING MODERN PRINCIPLES OF EXPERIMENTAL DESIGN AND ANALYSIS. THE GREATEST PRESENT NEED SEEMS TO BE TO APPLY SUCH PRINCIPLES TO THE TESTING OF HYPOTHESES DERIVED FROM CONSIDERATION OF PRIOR EMPIRICAL FINDINGS. STATUS STUDIES, SUCH AS COMPARING "FAMILIALS" WITH BRAIN-INJURED AMENTS ON THE BENDER GESTALT, DO ADD TO KNOWLEDGE AND THEORY, OF COURSE, BUT IT APPEARS TO THE INVESTIGATORS THAT MANIPULATION OF VARIABLES WITHIN THE CONTROL OF THE EXPERIMENTER SHOULD BE MORE FRUITFUL, PARTICULARLY WHERE THESE VARIABLES DERIVE FROM PRINCIPLES OF PSYCHOLOGY, PSYCHIATRY, MEDICINE, SOCIOLOGY, ANTHROPOLOGY, EDUCATION, AND OTHER RELATED DISCIPLINES. STUDY OF THE FEEBLEMINDED BY SOCIAL SCIENTISTS SEEMS LIKELY TO REMAIN CLOSER TO SERVICE THAN RESEARCH. PROBABLY A MORE OVER-ARCHING FRAME OF REFERENCE IS DESIRABLE, WITH DIMENSIONS OF INVESTIGATION CUTTING ACROSS VARIOUS LEVELS OF MENTAL DEVELOPMENT. THE RESEARCH CONCLUDED THAT MUCH FUNDAMENTAL (PURE) RESEARCH IS NEEDED IN ORDER TO DEVELOP THEORIES AND PRINCIPLES FROM WHICH RECOMMENDATIONS FOR PRACTICE CAN BE DEDUCED AND EVALUATED. (GC)

ED002777

REFERENCE
COPY

COOPERATIVE
RESEARCH
PROJECT

No. 149

Cooperative Research Project 6435:

"Restricted Generalization, Bias, and Loss of Power that May Result from Matching Groups"

University of Wisconsin
Dr. Julian C. Stanley

(149)

September 27, 1957

Restricted Generalization, Bias, and Loss of Power That May Result From Matching Groups[‡]

Julian C. Stanley and Ellen Y. Beeman

University of Wisconsin

Contents

		Page
1.	Introduction	2
2.	Randomized Groups	2
3.	Matched Groups	3
4.	The Analysis of Covariance	5
5.	Initially Intact Groups	6
6.	An Illustrative Analysis of Covariance	7
7.	Matched Pairs in Lieu of the Analysis of Covariance	10
8.	Discussion	10
9.	A Relatively Complex Design	12
10.	The Callahan-Keller Study	14
11.	Analysis of Covariance for Other Designs	17
12.	Group Matched for Means and Variances	19
13.	Matching in Recent Studies of Mental Deficiency	20
	a. Same Subjects, Measured More Than Once	21
	b. Matching Taken into Account	21
	c. Matching Ignored in Statistical Analysis	23
	d. No Matching	24
14.	Concluding Remarks	24
15.	References	26

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

Introduction

While surveying published experimental investigations in the area of mental deficiency, the authors found numerous studies where matching on the basis of antecedent variables such as MA, IQ, CA, and length of institutionalization and then ignoring in subsequent statistical analyses the effects of such matching seriously beclouded the interpretation of findings.

Matching seems virtually an occupational disease of researchers in this field; usually the analysis of covariance would be preferable. Perhaps covariance analysis is employed so seldom because there are few, if any, examples in the literature on mental deficiency to follow, though many statistics textbooks explain and illustrate this technique (e.g., 9, 12, 19, 23, 31, 41).

We shall tell why matching may be either suitable or unsuitable, depending upon the study, and by means of practical examples compare its results with those from analyses of covariance. We shall do this in considerable detail for two simple situations and then indicate that more complex designs may also be analyzed similarly.

Randomized Groups

When a researcher desires to apply one experimental treatment to a certain group and another experimental treatment (perhaps nothing or a placebo) to the other group, ideally he will draw N individuals randomly

---

[1]This study was facilitated by a research grant from the U. S. of Education, so this paper may be used freely for any purposes of the United States Government.

from the parent population to which he wishes to generalize statistically, assigning the first N/2 = n persons to a predesignated group and the remaining N/2= n to the other group. Because of this procedure, he knows that the two groups differ only by chance on any antecedent variable whatsoever, be it IQ, shoe size, or number of freckles.

Matched Groups

If his N is too small to give the significance test for the difference between means on the experimental (dependent) variable the power an experimenter desires (23, pp. 64-70), he may seek antecedent variables for all N subjects that, on the basis of prior experience are expected to correlate well with the dependent variable. Then, following tradition, he would probably match the two groups (a) individual by individual or (b) with respect to means and standard deviations, without special regard to individuals. If he does not want to discard individuals, thereby limiting the population to which he can generalize and perhaps also creating bias due to regression-toward-the-mean (29, 39), he may simply rank-order all the N scores on the matching variate and use a longitudinal split to assign subjects to one group or the other. Suppose, for example, that the intelligence-test scores of 20 third graders are as in Table 1. We assigned the first individual in a pair to the experimental group if a flipped coin fell "heads," and to the control group if it landed "tails." This is a "randomized block" design, each pair being a "block."

Pairing of this kind, without discarding subjects, will not restrict statistical inference. It may reduce the 'error' term considerably and thereby produce a more powerful significance test than would have resulted

Table 1

The Intelligence-Test Scores of 20 Third Graders as a Basis for Selecting Two Matched Groups

Individual	Pair	Score	Pair	Group	
				Experimental	Control
1	1	87	1	2	1
2		86	2	4	3
3	2	84	3	6	5
4		82	4	7	8
5	3	81	5	10	9
6		80	6	11	12
7	4	80	7	13	14
8		79	8	15	16
9	5	78	9	18	17
10		78	10	20	19
11	6	77			
12		77			
13	7	76			
14		76			
15	8	76			
16		73			
17	9	72			
18		72			
19	10	72			
20		69			

without pairing (23, pp. 83-87). In fact, an r of .75 between the matched groups on the dependent variable doubles the value of t. In general, at a given level of significance, $t_{(n-1\ d.f.)}$ approximately equals $t_{[2(n-1)\ d.f.]}\sqrt{1-r_{12}}$. See Stanley (35).

The Analysis of Covariance

Instead of pairing scores on the basis of the matching variable and then testing the null hypothesis that the mean of the dependent-variate differences between pairs in the normally distributed population of differences from which this sample was drawn randomly is 0, one might instead select the groups randomly, treat the matching variable (intelligence-test score, in our example) as a covariable, and employ the analysis of covariance. If the regression of the dependent variable on the covariable is linear, this prodedure will often be superior to the matching procedures mentioned above. Also, it easily accommodates unequal numbers in the experimental and control groups.

Cox (10) discusses "blocking" vs. the analysis of covariance and concludes that "the methods based on covariance are preferable to the simpler methods based on blocking only if the correlation coefficient between y and x is at least 0.6 . . ." (p. 157). This result applies to situations like the one depicted in our Table 1, where no subjects are discarded and there is a manipulable experimental (dependent) variable, the researcher's only interest in the concomitant variable being to increase the precision of this experiment by reducing the magnitude of the error term. Where intact groups are concerned and/or subjects are discarded nonrandomly, on the basis of the antecedent variable (s), Cox's

conclusion is not directly relevant.

Initially Intact Groups

For many studies, a more important use of the analysis of covariance than for increasing power is statistical "equating" of non-random groups, where random selection was not feasible. As an example, suppose that we want to compare reading test scores of boys and girls in the third grade. If intelligence test and reading raw scores are as shown in Table 2, it is obvious that the boys have a somewhat higher mean score than the

Table 2

Intelligence-Test and Reading Scores of 10 Boys and 10 Girls

Boys (b)		Girls (g)	
Intelligence-Test Score (X_b)	Reading Score (Y_b)	Intelligence-Test Score (X_g)	Reading Score (Y_g)
79	61	72	78
82	82	87	75
80	78	78	81
77	68	73	81
81	62	72	68
80	74	78	72
86	69	76	72
72	59	69	71
84	60	76	80
77	73	76	81
Means 79.8	68.6	75.7	75.9

girls on the intelligence test (79.8 vs. 75.7), probably due largely to the boys' being slightly older. Do third-grade boys and girls in the hyperpopulations of which these scores may logically be considered random samples differ significantly in mean reading score when the mean intelligence-test scores are equalized statistically by the analysis of covariance? Let us use our undesirably small but computationally convenient set of scores from Table 2 to illustrate this analysis. See Table 3. In carrying out the analysis of covariance, we are implicitly performing the two analyses of variance—of intelligence test scores and of reading scores—and then combining them in a certain way if justifiable. Table 3 contains the sums of squared deviations from the means for intelligence scores (X) and reading scores (Y) and the sum of crossproducts (XY) expressed in deviation form.

Intelligence scores for the two groups do not differ significantly in variability, nor do reading scores; the respective F's, each with 9,9 d.f., are 1.56 and 2.69. Of course, because of the small samples involved in this illustration, these F's do not have much power; errors of the second kind are probably quite large, making it unlikely that we would refute the null hypothesis even if it were false. But Box (5) shows that when the groups are of equal size, as here, the analysis of variance is little affected by moderate inequality of variances.

The boys do not quite exceed the girls significantly at the .05 level on mean intelligence-test score: $F_{1,18} = 4.23$.

The girls exceed the boys significantly at the .025 level on mean reading score: $F_{1,18} = 5.99$.

An Illustrative Analysis of Covariance

If (a) the regression of reading scores on intelligence scores is

Table 3

Analyses of Variance and Covariance of Scores from Table 2

Source of Variation	d.f.	Sums of Squares and Products Σx^2	Σxy	Σy^2	b_{yx}	r_{xy}	Errors of Estimate Sum of Squares	d.f.
Total	19	441.75	- 53.75	1067.75	- 0.1217	- .0783	1061.2100	18
(Within Boys)	(9)	(139.60)	(61.20)	(584.40)	(0.4384)	(.2143)	(557.5702)	(8)
(Within Girls)	(9)	(218.10)	(34.70)	(216.90)	(0.1591)	(.1596)	(211.3792)	(8)
							(768.9494)	(16)
Pooled Within Sexes	18	357.70	95.90	801.30	0.2681	.1791	775.5890	17
Between Sexes	1	84.05	- 149.65	266.45	- 1.7805	- 1.0000	285.6210 (By subtraction)	1

$x^F_{9,9} = 218.10/139.60 = 1.56$ $P > .05$

$y^F_{9,9} = 584.40/216.90 = 2.69$ $P > .05$

} Variances not significantly heterogeneous.

$x^F_{1,18} = \frac{84.05/1}{357.70/18} = 4.23$ $.95F_{1,18} = 4.41$ Sexes do not quite differ significantly on intelligence test scores.

$y^F_{1,18} = \frac{266.45/1}{801.30/18} = 5.99$ $.975F_{1,18} = 5.98$ Girls exceed boys significantly on reading scores.

$F_{1,16} = \frac{(775.5890 - 768.9494)/1}{768.9494/16} < 1$ Regressions not significantly heterogeneous.

$F_{1,17} = \frac{(801.3000 - 775.5890)/1}{775.5890/17} < 1$ Reduction in sums of squares for error not significant.

$y^F_{1,17} = \frac{285.6210/1}{775.5890/17} = 6.26$ $.975F_{1,17} = 6.04$ Girls exceed boys significantly on reading scores.

linear and (b) this regression does not differ significantly for boys and girls, then we can use intelligence scores efficiently as a covariable. We do not have a large enough group of third graders to detect more than the grossest departures from linearity, so since relationships between tests of intelligence and reading are usually linear (38, p. 184), we shall dispense with significance tests for linearity of regression and merely call the reader's attention to McNemar's (23, p. 268-275) lucid explanation of the procedure.

The regression coefficient (slope) of reading on intelligence, b_{yx}, is 0.44 for the boys and 0.16 for the girls. Do these two slopes differ significantly? If they do, we have intelligence interacting with sex and cannot legitimately pool the sums of crossproducts for boys and girls. Table 3 shows the appropriate significance test for the difference between the two regression coefficients (31, Table 12.7): $F_{1,16} = \frac{6.6396/1}{768.9494/16} < 1$, clearly not significant at the .05 level.

Now we are ready for the analysis of covariance, also shown in Table 3. Compare the $F_{1,17} = 6.26$ with the analysis-of-variance $F_{1,18} = 5.99$. This rise from 5.99 to 6.26 does not reflect an important increase in precision because the reduction in the 'error' sum of squares from 801.3000 to 775.5890 in Table 3 is not significant: $F_{1,17} < 1$. In fact, $\frac{801.3000}{18} < \frac{775.5890}{17}$, so the degree of freedom was sacrificed in vain.

On the other hand, the sum of squares between sexes increased from 266.450 to 285.621. Therefore, both the numerator and the denominator of the analysis-of-variance ratio are a little smaller than the numerator and the denominator of the analysis of covariance ratio: $266.45/44.52 = 5.99$ versus $285.621/45.62 = 6.26$. Thus we gained a little power by adjusting

means but lost a bit in unsuccessfully trying to reduce the size of the error term by taking into account the within-sex r of .18.

The original reading-test means were 68.6 for boys and 75.9 for girls, a difference of 7.3. The adjusted means, which take into account the original imbalance on intelligence and the regression of reading on intelligence, are $68.6 - 0.2681(79.80 - 77.75) = 68.05$ for the boys and $75.9 - 0.2681(75.70 - 77.75) = 76.45$ for the girls, a difference of 8.4.

Matched Pairs in Lieu of the Analysis of Covariance

Instead of performing an analysis of covariance on the scores in Table 2, we can match boy-girl pairs as closely as possible with respect to intelligence score and then carry out an analysis of variance on the resulting 10 pairs of reading scores. See Table 4 for this procedure.

The results are quite disappointing because the correlation between reading scores of the pairs is $(866.1 - 584.4 - 216.9)/ 2\sqrt{(584.4)(216.9} = -.09$, which actually increases the 'error' term slightly and hence makes $F_{1,9} = 5.54$; contrast this with the analysis-of-covariance $F_{1,17} = 6.26$. Our attempt to induce positive correlation between reading scores by pairing on mental-test scores was not fruitful in this illustration; in other instances it might be, of course.

Discussion

Lindquist (19, pp. 317-339) presents a thorough discussion of the analysis of covariance, showing assumptions underlying it. One of the most serious of these is that the covariable should be the only relevant factor on which the groups differ significantly initially. In our illustration above, we would need to assume that the boys did not differ

Table 4

Reading Scores of Paired Boys and Girls; Scores from Table 2

Pair Number	Boys	Girls	Girls - Boys
1	69	75	6
2	60	81	21
3	82	72	- 10
4	62	72	10
5	78	80	2
6	74	81	7
7	61	81	20
8	68	78	10
9	73	68	- 5
10	59	71	12
Means	68.6	75.9	7.3

$$F_{1,9} = \frac{(7.3 - 0)^2}{866.1/(9)(10)} = 5.54 \qquad .025 < P < .05$$

significantly from the girls initially on any factor other than mental age, such as socio-economic status or distance to nearest library, that was correlated significantly with reading score. For youngsters selected from the same school, this assumption might be reasonable. If, however, we took boys from a public school and girls from a nearby parochial school, we might not be able to generalize logically to boys vs. girls in public school or boys vs. girls in parochial schools, even though mental age were 'controlled' statistically.

This danger is especially great when one intact group is employed as

the "experimental group" and the other intact group as the "control group," usually because of their convenient separate locations. We would hardly consider using brain-damaged youngsters as the experimental group in a learning study and "familial" feebleminded youngsters as controls, no matter how many relevant covariables we could secure. However, educators have not hesitated in the past to try out experimental procedures in laboratory schools, using matched public-school younsters as controls. Conclusions from such studies will usually be equivocal, for we cannot equate the two groups for some factors, such as willingness of parents to send a child to a laboratory school, that resulted initially in the two different kinds of groups. If these factors are related to the dependent variable (outcome of the experiment), then the comparison will remain biased--usually in favor of the laboratory school, at least where average or above average students are concerned.

Two solutions to this dilemma of not being able to secure all necessary covariables are: (a) perform your experiment with a random half of the laboratory schools as experimental schools and the other half as controls, and similarly for public schools; or (b) have randomly assigned experimental and control sections within each school. It may also be possible to obtain covariables for either of these situations in order to get increased precision and to introduce explicit factors such as sex and socio-economic level into the experimental design itself.

A Relatively Complex Design

As an illustration of efficient design, consider a study now under way that involves comparing with standard control classes instruction in high-school physics carried out largely via well prepared sound films.

After carefully considering relevant factors, the designers of this investigation chose mean score of school on a certain intelligence test (above average vs. below average), semester hours of physics credit the instructor has (high, average, low), score of instructor on a standardized physics test (above average, below average), and teaching method (film vs. usual procedure) as their four explicit independent variables, yielding $2 \times 3 \times 2 \times 2 = 24$ cells, in each of which there were three schools. Only teaching method could be assigned randomly, within each of the 12 "status" cells; the other three factors were indigenous to the schools or instructors and were not subject to manipulation in this study.

Further antecedent variables possibly related to mean physics achievement of a school at the end of the year are available: size of class, size of school, and per pupil expenditure on laboratory supplies. These may be used as covariables in the above design. Degrees of freedom for the "within" mean square would be $24(3 - 1) = 48$ without the covariables. Each covariable reduces the error d.f. by 1, so although computational labor will be increased considerably by introducing one or more covariables into a four-factor design, the loss of degrees of freedom is slight and the extra effort may be worth while. (After all, when one devotes a whole year to designing and carrying out an experiment, should he shy away from a few extra days of statistical analysis?) See Lindquist (19, pp. 335-336) and McNemar (23, pp. 354-355) for the computational rationale.[2]

---

[2]Some transformations of the covariates and dependent variate (mean score of each class on a physics test at the end of the year) may be essential in order to secure linear regression, independence of means and variances, and additivity of main effects.

The Callahan-Keller Study

Callahan and Keller (7) have generously provided us data from their study of the relationship between high and low digit-span and scores on the Children's Anxiety Pictures (CAP). In somewhat rearranged form, the design is shown in Table 5. We have analyzed these scores in three ways:

Table 5

The Callahan-Keller (7) Data

Race (r)	Low Digit Span		High Digit Span	
	IQ (X_1)	CAP Score	IQ (X_h)	CAP Score
White	71	39	70	24
	63	34	62	13
	68	23	67	12
	54	27	54	21
	79	16	80	21
	85	34	84	9
	68	26	64	18
Negro	66	25	66	28
	82	37	82	16
	74	27	75	16

(a) as matched pairs by the analysis of variance; (b) by the analysis of variance, with pairing ignored; (c) by the analysis of covariance, with race ignored. For their data, (a) is most defensible, since they actually did match subjects, probably discarding many, but we may gain some insight into what might have happened had they chosen to use IQ as a covariable

instead of matching on it. (Note that we have made white vs. Negro an explicit factor in the design instead of a matching factor. The small range in age, another matching variable, can probably be ignored safely in our analyses).

The first analysis, shown in Table 6, yields an $F_{1,8} = 10.87$ for

Table 6

Analysis of Variance of CAP Scores from Table 5

Source of Variation	d.f.	Σy^2	Mean Square
Race (r)	1	20.1524	20.1524
Between individuals within race (b)	8	286.0476	35.7560
Digit span (d)	1	605.0000	605.0000
r x d	1	3.8095	3.8095
Individuals x d within r	8	445.1904	55.6488
Total	19	1360.2000	—

$$F_{1,8} = \frac{605.0000}{55.6488} = 10.87 \qquad .01 < P < .025$$

the digit span factor, significant beyond the .025 level with a two-tailed test. [Callahan and Keller (7) themselves used a simple one-tailed t-test, since they had an explicit directional hypothesis. See Kimmel (17) for a brief summary of the controversy concerning one-tailed vs. two-tailed tests. For our purposes, we shall compare the F's themselves.]

Callahan and Keller, ignoring race and matching and utilizing one more case

(unpaired) than shown in our Table 5, obtained an $F_{1,19}$ of about 16.8, which is roughly comparable to our $F_{1,18}$ of 14.42 in Table 8.

Note that the $MS_b = 35.7560$ of Table 6 is less than the $MS_{(i \times d)} = 55.6488$ there, indicating that the pooled within-race correlation of scores on the dependent variable (CAP) for pairs is negative. An estimate of its magnitude (36) is given by $\frac{35.7560 - 55.6488}{35.7560 + 55.6488} = -.22$. Therefore, Callahan and Keller did not succeed in producing positive correlation between subjects paired on the basis of IQ; consequently, they disregarded the matching in their analysis.

Table 7 shows that if we ignore matching but retain race as an

Table 7

Matching Disregarded

Source of Variation	d.f.	Σy^2	Mean Square
Race (r)	1	20.1524	20.1524
Digit span (d)	1	605.0000	605.0000
r x d	1	3.8095	3.8095
Within (w)	16	731.2380	45.7024
Total	19	1360.2000	—

$$F_{1,16} = \frac{605.0000}{45.7024} = 13.24 \qquad .001 < P < .005$$

explicit factor, the magnitude of F—now with 16 d.f. for error—rises to 13.24 because by eliminating the influence of the negative correlation we decrease the mean square for error from 55.6488 to 45.7024. In other

words, we pool SS_b and $SS_{(i \times d)}$ of Table 6 to yield the SS_w of Table 7. This is a "sometimes-pool" procedure (6), presumably justified because $F_{8,8} = \frac{55.6488}{35.7560}$ if far from being significant.

Finally, in Table 8 we show the analysis of covariance of Callahan and Keller's data, ignoring race. This, too, is nearly as disappointing as the outcome of the analyses in Tables 6 and 7, because the "adjusted" mean square for error with 17 d.f. is larger than the unadjusted mean square with 18 d.f. (44.2457 vs. 41.9556), making $(F_{1,17} = 13.72) < (F_{1,18} = 14.42)$. Either figure is highly significant, of course, but it is obvious that the effort of matching subjects on IQ was wasted and that any biases resulting from this matching might have been avoided without lowering the precision of the study. Indeed, if the time needed for matching had been employed to secure more subjects, a gain in precision would undoubtedly have occurred. In a study where, unlike this one, the factor studied is not very potent, the increased precision may be needed to demonstrate a significant difference.

Let us emphasize that we have analyzed the Callahan-Keller scores merely as a convenient example of a situation in which matching (or covariance analysis) was ineffective. Their conclusions remain unaltered by our further manipulations.

Analysis of Covariance for Other Designs

Our two detailed illustrations for the analysis of coveriance show how to handle just the one-way classification with a single covariable; there may be more than two groups, and they need not be of equal size, though in our examples they are.

The analysis of covariance may be applied to higher-order classifica-

Table 8

Analysis of Covariance, Ignoring Race

Source of Variation	d.f.	Sums of Squares and Products			b_{yx}	r_{xy}	Errors of Estimate	
		Σx^2	Σxy	Σy^2			Sum of Squares	d.f.
Total	19	1632.2000	- 37.2000	1360.2000	- 0.0228	- .0250	1359.3522	18
(Low DS)	(9)	(786.0000)	(89.0000)	(451.6000)	(0.1132)	(.1494)	(441.5224)	(8)
(High DS)	(9)	(844.4000)	(- 159.2000)	(303.6000)	(- 0.1885)	(- .3144)	(273.5850)	(8)
Within	18	1630.4000	- 70.2000	755.2000	- 0.0431	- .0633	752.1774	17
Digit Span	1	1.8000	33.0000	605.0000	18.3333	1.0000	607.1748	1

$$y F_{1,18} = \frac{605.0/1}{775.2/18} = 14.42 \qquad .001 < P < .005.$$

$$F_{1,16} = \frac{(752.1774 - 441.5224 - 273.5850)/1}{(441.5224 + 273.5850)/16} < 1.$$

Therefore, the two regressions are not significantly heterogeneous.

$$y F_{1,17} = \frac{607.1748/1}{752.1774/17} = 13.72 \qquad .001 < P < .005.$$

tions such as poly-factor designs, randomized blocks, latin squares, and split-plot designs (e.g., 1, 16, 31, 40). Basic principles of Tables 3 and 8, with appropriate extensions, apply to these situations, also.

Groups Matched for Means and Variances

Sometimes the experimenter juggles his subjects on the basis of an antecedent variable such as mental age and by discarding subjects here and there manages to equate the means and variances of the various groups on the antecedent variable without explicitly matching subjects or necessarily having the same number of subjects in each group. Usually this procedure restricts and makes ambiguous the populations to which generalizations can be made, in addition to causing a regression-toward-the-mean effect (29, 39) that biases the outcome of the experiment in favor of the groups initially (before matching) superior on the antecedent variable. It would be better, therefore, not to eliminate subjects, but instead to treat the antecedent variable as a covariable and to apply the analysis of covariance to the dependent variate.

When, however, an investigator has already matched his groups, he must at least take this matching into account in his statistical analysis; otherwise, if the antecedent variable is linearly related to the dependent variable, he may accentuate errors of the second kind (23) by rejecting the null hypothesis too seldom when in fact it is false. Edwards (13, pp. 282-288) shows that for two matched groups the analysis of covariance is the proper procedure (though he assumes that the group means on the antecedent variable are precisely equated); it would also be used with more than two groups and with more than one antecedent variable. Also see McNemar (23, pp. 363-366).

Let us emphasize, however, that the analysis of covariance applied to groups equated by discarding certain aberrant subjects does not compensate for restricted generalization and for bias due to regression effects. Usually, it would be preferable to take a random sample from each of the populations concerned and then equate means and increase precision statistically via the analysis of covariance, without any matching whatsoever.

Researchers in the area of mental deficiency seem to regard matching of some sort as the sine qua non of experimentation. As we have already mentioned, Cox (10) points out certain conditions under which "blocking" is preferable to the analysis of covariance--especially when assumptions of homogeneity of regressions and linearity are violated--but he is not dealing with the situation in which certain experimental units are discarded from the design in a non-random manner. Generally, then, investigators in this field will find the analysis of covariance preferable to matching. Of course, there are some situations, such as the typical educational 'experiment' of the 1930's in which an experimental method was applied to students in "laboratory" schools, and students in public schools were used as controls, where the basic design itself is faulty and no kind of statistical procedure will shed much light on the proposed hypothesis.

Matching in Recent Studies of Mental Deficiency

Though most textbooks on statistics emphasize that significance tests for "correlated" means differ from those for "uncorrelated" means (23, pp. 108-111), frequently investigators in the field of mental deficiency fail to take into account such constraint on sampling fluctuations unless they match subjects instead of groups. The literature abounds with

illustrations of this oversight, which leads to loss of power if matching was effective in producing correlation between groups on the dependent variable. We shall consider selected examples of correct and incorrect practice, in so far as the published study permits us to determine whether or not the error term was adjusted to allow for the influence of matching.

Same Subjects, Measured More Than Once

Atchison (2) compared the WISC verbal and performance IQ's of 80 mentally defective Negro children, properly using the r of .39 in computing his CR (although he does not state in the article what method was employed, the CR we computed for "correlated" means agrees with the one he reports).

Likewise, Rosenblum, Keller, and Papania (28) seem to have taken correlation into account in their analysis of mean scores secured by 30 mentally deficient boys on the Davis-Eells "Games," Stanford-Binet, WISC, and California Test of Mental Maturity.[3]

Matching Taken into Account

Halpin (15, p. 485) "matched by pairs on mental age, chronological age and intelligence" 15 brain-injured children with 15 non-brain-injured children. We can be rather sure that she tested the differences on the dependent variate appropriately because her Table II shows t's with 14 d.f. each, rather than the 28 d.f. that would have resulted had she ignored correlation between pairs and used a pooled within-group error

[3]Dunnett's procedure (11, 34), with a one-tailed significance test, might lend itself well to their problem of comparing several tests with the "Games," although it was not available at the time their study was done.

term. It is unlikely, however, that she utilized the three antecedent variables (MA, CA, and IQ) optimally in her pairing, and it is probable that by matching rather closely she was forced to discard a considerable number of subjects from the study, thereby restricting her generalization of these findings in an indeterminate manner. If the two populations were different with respect to mean CA, MA, and IQ (as very likely they were if she found it desirable to match subjects), regression-toward-the-mean bias probably operated in the results, perhaps even to produce a seemingly significant difference (29, 39). This was a study where the analysis of covariance might well have been used in lieu of matching. With both MA and CA covariables, Halpin would have had 26 d.f. for error instead of 14, even with only 30 subjects.

Bensberg (4) paired a "familial subject" with a brain-injured subject having about the same MA and CA. (These seem to be the favorite antecedent variables of researchers in mental deficiency.) He had 161 such pairs, 73 of them female and 88 male. His reported degrees of freedom of 72 and 87, respectively, for t tests reveal that he considered correlation by dealing with differences. He might, however, have tested the significance of the interaction of sex with the familial vs. brain damage factor, though there was an experimental complication that may have precluded this.

In his Table 1, Bensberg (4) carried out a more complex analysis that puzzles the writers, who plan to correspond with him about it. It appears that, though he planned to keep the matching in mind here, the error sum of squares consists of four pooled sums of squares (three interactions involving "pairs," p the between-pairs sum of squares) that may not be homogeneous. Also, his degrees of feedom seem excessive; 33 pairs, 2 etiologies, and 8 figures should yield only (33)(2)(8) - 1 =

527 d.f., not the total of 593 that Bensberg shows. There may be some relevant point in the short report that we are overlooking which would justify these discrepancies, but thus far we have not found it, unless the explanation is simply that Bensberg treated 8 figures as if they had 8 d.f. in all appearances, rather than just 7.

Goldberg (14) matched all three groups for CA and two of them for intelligence. He states that "all the Fisher *t* ratios were based on the formula for finding the *t* value between the means of correlated groups" (p. 550). A portion of his analysis perplexes us, especially when in Table I he reports both *t*'s and *F*'s for the same three possible comparisons among means of the three groups, with the latter smaller than the former and having different levels of significance. Except for the caption of the table, no reference to these *F*'s seems to appear in the article. For the simple comparisons that he makes, t^2 should equal *F*. We surmise that his *F*'s do not take correlation into account, while the *t*'s do. If so, the *F*'s are unnecessary and somewhat misleading.

Matching Ignored in Statistical Analysis

Examples of matching on antecedent variables such as CA, MA, IQ, years of institutionalization, diagnosis, social status, and sex, apparently without taking the correlation thereby engendered between groups into account, are numerous (e.g., 3, 20, 22, 24, 25, 27, 30, 32, 42). The influence of this oversight on conclusions reached may vary from negligible to overwhelming, depending upon the degree of correlation and the potency of the experimental variable.

Curiously enough, sex is seldom made an explicit dichotomous factor in an experiment so that its influence and interaction(s) can be studied; instead, it often appears as a matching variable. We suspect that the

typical researcher in this area shuns designs having a second factor because of (a) ignorance about their advantages and/or (b) fear that he will have trouble analyzing his results.

No Matching

Not all investigators insisted upon matching subjects. So far as we could determine, no researcher in the field of mental deficiency has published a study in which he utilized the analysis of covariance, but some did employ other techniques, especially the analysis of variance, without antecedent matching (e.g., 8, 18, 21, 26, 33, 37).

Concluding Remarks

A few researchers in mental deficiency are performing theoretically oriented experiments, utilizing modern principles of experimental design and analysis. The greatest present need seems to be to apply such principles to the testing of hypotheses derived from consideration of prior empirical findings. Status studies, such as comparing "familials" with brain-injured aments on the Bender Gestalt, do add to knowledge and theory, of course, but it appears to the writers that the explicit manipulation of variables within the control of the experimenter should be more fruitful, particularly where these variables derive from principles of psychology, psychiatry, medicine, sociology, anthropology, education, and other related disciplines.

Study of the feebleminded qua feebleminded by social scientists seems likely to remain closer to service than to research. Probably a more over-arching frame of reference is desirable, with dimensions of investigation cutting across various levels of mental development. Some workers in mental deficiency already realize this fully and couch their studies in terms of reward, level of aspiration, and other concepts,

rather than merely feebleminded vs. normal intelligence. It is tempting to attack mental deficiency per se, the way the lay public mistakenly supposes that the biochemist seeks a cure for cancer, but our feeling is that much fundamental ("pure") research is needed in order to develop theories and principles from which recommendations for practice can be deduced and evaluated.

References

1. Anderson, Richard L., & Bancroft, Theodore A. Statistical theory in research. New York: McGraw-Hill, 1952.

2. Atchison, Calvin O. Use of the Wechsler Intelligence Scale for Children with eighty mentally defective Negro children. Amer. J. Ment. Defic., 1955, 60, 378-379.

3. Bensberg, Gerard J. A test for differentiating endogenous and exogenous mental defectives. Amer. J. Ment. Defic., 1950, 54, 502-506.

4. Bensberg, Gerard J. Performance of brain-injured and familial mental defectives on the Bender Gestalt Test. J. Consult. Psychol., 1952, 16, 61-64.

5. Box, George E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one-way classification. Ann. math. Statist., 1954, 25, 290-302.

6. Bozivich, Helen; Bancroft, Theodore A.; and Hartley, Herman O. Power of analysis of variance test procedures for certain incompletely specified models. Ann. math. Statist., 1956, 27, 1017-1043.

7. Callahan, Roger J., & Keller, James E. Digit span and anxiety: An experimental group revisited. Amer. J. ment. Defic., 1957, 61, 581-582.

8. Cantor, Gordon N., & Hottel, John V. Discrimination learning in mental defectives as a function of magnitude of food reward and intelligence level. Amer. J. ment. Defic., 1955, 60, 380-384.

9. Cochran, William G., & Cox, Gertrude M. Experimental designs. (2nd Ed.) New York: Wiley, 1957.

10. Cox, David R. "The use of a concomitant variable in selecting an experimental design." Biometrika, 1957, 44, 150-158.

11. Dunnett, Charles W. A multiple comparison procedure for comparing several treatments with a control. J. Amer. statist. Assn., 1955, 50, 1096-1121.

12. Edwards, Allen L. Experimental design in psychological research. New York: Rinehart, 1950.

13. Edwards, Allen L. Statistical methods for the behavioral sciences. New York: Rinehart, 1954.

14. Goldberg, Franklin H. The performance of schizophrenic, retarded, and normal children on the Bender-Gestalt Test. Amer. J. ment. Defic., 1957, 61, 548-555.

15. Halpin, Virginia G. Rotation errors made by brain-injured and familial children on two visual-motor tests. Amer. J. ment. Defic., 1955, 59, 485-489.

16. Kempthorne, Oscar. The design and analysis of experiments. New York: Wiley, 1952.

17. Kimmel, Herbert D. Three criteria for the use of one-tailed tests. Psychol. Bull., 1957, 54, 351-353.

18. Korstvedt, Arne; Stacey, Chalmers L.; & Reynolds, William F. Concept formation of normal and subnormal adolescents on a modification of the Weigl-Goldstein-Scheerer Color Form Sorting Test. J. clin. Psychol., 1954, 10, 88-90

19. Lindquist, Everet F. Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin, 1953.

20. Lombard, J. P.; Gilbert, Jeanne G.; & Donofrio, Anthony F. The effects of glutamic acid upon the intelligence, social maturity and

adjustment of a group of mentally retarded children. Amer. J. ment. Defic., 1955, 60, 122-132.

21. Lubin, Nathan M. The effect of color in the TAT on productions of mentally retarded subjects. Amer. J. ment. Defic., 1955, 60, 366-370.

22. McCulloch, Thomas L.; Reswick, Joseph; & Roy, Irving. Studies of word learning in mental defectives. I. Effects of mental level and age. Amer. J. ment. Defic., 1955, 60, 133-143.

23. McNemar, Quinn. Psychological statistics. (2nd ed.) New York: Wiley, 1955.

24. Mecham, Merlin J. The development and application of procedures for measuring speech improvement in mentally defective children. Amer. J. ment. Defic., 1955, 60, 301-306.

25. Murphy, Mary M. Comparison of developmental patterns of three diagnostic groups of middle grade and low grade mental defectives. Amer. J. ment. Defic., 1956, 61, 164-169.

26. Reynolds, William F., & Stacey, Chalmers L. A comparison of normals and subnormals in mirror drawing. J. genet. Psychol., 1955, 87, 301-308.

27. Rosenblum, Sidney. Ethnocentrism and intolerance of ambiguity in mentally retarded children. Amer. J. ment. Defic., 1957, 61, 567.573.

28. Rosenblum, Sidney; Keller, James E.; & Papania, Ned. Davis-Eells ("culture-fair") test performance of lower-class retarded children. J. consult. Psychol., 1955, 19, 51-54.

29. Rulon, Phillip J. Problems of regression. Harv. educ. Rev., 1941 11, 213-223.

30. Shaw, Marvin E., & Bensberg, Gerard J. Level of aspiration phenomena in mentally deficient persons. J. Pers., 1955, 24, 134-144.

31. Snedecor, George W. Statistical methods applied to experiments in agriculture and biology. (5th Ed.) Ames, Iowa: Iowa State College Press, 1956.

32. Solomon, Paul. A note on rigidity and length of institutionalization. J. clin. Psychol., 1954, 10, 391-392.

33. Solomon, Paul. Differential Rorschach scores of successfully and unsuccessfully placed mental defectives. J. clin. Psychol., 1955, 11, 394-397.

34. Stanley, Julian C. Additional "post-mortem" tests of experimental comparisons. Psychol. Bull., 1957, 54, 128-130.

35. Stanley, Julian C. Improper t-tests, with special reference to the Diller-Beechley note. J. clin. Psychol., 1953, 9, 391-393.

36. Stanley, Julian C. K-R_{20} as the mean intercorrelation among items. 14th Yearbook of the National Committee on Measurements Used in Education. Princeton, N. J.: Dr. Edith Huddleston, Educational Testing Service, Editor, 1957 (in press).

37. Swanson, Robert. Perception of simultaneous tactual stimulation in defective and normal children. Amer. J. ment. Defic., 1957, 61, 743-752.

38. Thorndike, Robert L. Personnel selection. New York: Wiley, 1949.

39. Thorndike, Robert L. Regression fallacies in the matched-group experiment. Psychometrika, 1942, 7, 85-102.

40. Truitt, Jeanne T., & Smith, H. Fairfield. Adjustment by covariance and consequent tests of significance in split-plot experiments. Biometrics, 1956, 12, 23-39.

41. Walker, Helen M., & Lev, Joseph. Statistical inference. New York: Holt, 1953.

42. Waters, Thomas J. Qualitative vocabulary responses in three etiologies of mental defectives. Train Sch. Bull., 1956, 53, 151-156.